An Analysis Of Census Dataset And Bank Dataset Using Classification Algorithms

Sangavi N B.Vinothini Dr. Premalatha K

Abstract: data mining is a process of extracting valuable information from large set databases. Classification a supervised technique is assigning data samples to target classes. In this system, it uses various classification algorithms namely decision trees, SVM, random forest and neural network in two datasets census and banking datasets. This system will classify and analyses the best suited algorithm which gives maximum accuracy among the other algorithms for both the datasets. The accuracy in these algorithms has been calculated by sensitivity and specificity. Evaluation of these models has been calculated by the error rate with respect to the classes. It uses census dataset and finds whether the income above 50k or below 50k. Bank Dataset finds whether the subscribe the loan or not. Error matrix consists of true positive, false positive, true negative and false negative values. The analysis of algorithm which finds the better algorithm with respect to the accuracy, error rate and efficiency with respet to these datasets.

Keywords: Decision tree, SVM, Random forest model, neural network model

1. INTRODUCTION

Census dataset contains various attributes such as age, work class, education, final weight, marital status, occupation, relationship, sex, capital gain and capital loss. Bank Marketing Dataset (BMD) contains eighteen attributes mainly job, maritial status, housing loan and home loan etc. The target is to find the whether they subscribe a loan or not. Data mining uses many tools to find out the behavior of the models. The various models has been used are SVM, random forest, decision tree and neural network. Random forests uses supervised learning method used for classification. Random forest model allows to construction of many decision trees with respect to the variables. Decision tree is used to solve the problem using both regression and classification. Each internal node is an attribute and each leaf node is a class label. SVM is a supervised learning which sorts two classes using the hyper plane. Neural network is a thousands of nodes are interconnected used for sales forecasting and time series prediction. There are three layers such as hidden layer, input layer and output layer[8].

2. OBJECTIVE

The main objective of the project is to find the performance efficiency and accuracy of the algorithms and to know which best suited one among those algorithms in two datasets. The prediction over these algorithms has been calculated with respect to specificity, sensitivity and accuracy. The various models has been considered such as random forest, decision tree, SVM and neural network.

3. PERFORMANCE MEASURE

i) Accuracy

Accuracy is defined as number of correct assessments by the total number of assessments. It is calculated by the specificity and sensitivity values[1].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is True Positive values, TN is True Negative, FP is False Negative and FN is False Negative values

ii) Sensitivity

Sensitivity is calculated from number of true positive values by number of all negative assessments. Sensitivity is observed by the proposition of positive values divided by the no of observation. It is calculated by true positive and false negative [1].

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP is True Positive values, FN is False Negative values

iii) Specificity

Specificity is calculated from number of true negative assesements by the total number of negative assesments. Specificity is calculated from true negative and false positive values[1].

$$Specificity = \frac{TN}{TN + FP}$$

where TN is True Negative and FP is False Negative values.

4. METHODOLOGIES

The census dataset and banking dataset has been chosen for the analysis of data. It is taken from UCI machine learning repository [7]. The attributes of the census dataset are age, work class, occupation, capital gain, capital loss, education, marital staus, race, sex, relationship, education num, hours of work and native country. Age, capital gain, capital loss, hours of work are continuous valued attribute whereas work class, occupation, education, marital staus, race, sex, relationship are categorical valued attributes. The classes of census dataset is above 50k or below 50k [2].

4.1 DECISION TREE

Decision tree model allows to classify the root nodes and leaf nodes. Root node will classify the nodes and allow to target the nodes. The leaf node are internal nodes which are attributes of the dataset. The root nodes are the classes which target either 0 or 1. The conditional probability will be checked with respect to the nodes. This model gives the tree representation with respect to the attributes. It is used to represent the root nodes as final classes and other internal nodes are attributes of classes. The relative error and standard error are noted and it will identify the root node. It

allows to identify the root node and root node error. Rattle allows to execute the model and time taken to execute the model is noted. Relationship is considered as root node in this dataset[3].

4.2 RANDOM FOREST

Random forest algorithm is a combination of many decision trees. It is a supervised learning technique that allows to build the forest with maximum of many decision trees. The implementation over the decision tree is used in random forests. The decision tree implementation is simple when compared to the random forest. It builds the model and indicates the error and time taken to execute the model. It will not allow categorical variables more than 32 levels. In our dataset, the attribute country has more than 32 categorical values [3].

4.3 SVM

SVM is linear regression used to classify the classes using hyper plane. This model uses line or hyper plane for classification of two sets using training dataset. It is a supervised learning technique that separate the two classes by the hyper plane. It considers some parameters for tuning. They are kernel, gamma, regularization and margin. Misclassification can be avoided by the tuning parameters in SVM. Smaller margin will be optimized easiliy by the regularization. Kernel allows to solve the equation by the support vector machine. The error of the model is given as relative error. This also allows to know the time taken to complete the execution the model[5].

4.4 NEURAL NETWORK MODEL

Neural Network model is a model which has three layers. Initially input is given to the input layer, then to the hidden layer and then to the output layer. Output layer is drawn from the hidden layer. It is a bit complex when comparing with other models. Neural network model is derived from biology i.e neurons in our brain. It process the information in parallel and is done with how neurons will work in brain. It is easy for humans to process information in brain. But it is difficult to formulate the information in the brain. One application of neural network is optical character recognition to object detection. It allows the brain that transform the information [5].

5. EVALUATION

5.1 ERROR MATRIX CALCULATION

It should calculate the accuracy, specificity and sensitivity of the models by error matrix of both the datasets

i) Decision tree:

The overall error rate of decision tree in census and banking dataset is 16.3% and 9.74%

```
Error matrix for the Decision Tree model on adult1.csv [test] (counts):

       Predicted
Actual  <=50K >50K Error
  <=50K  3502  217   5.8
  >50K    579  587  49.7

Error matrix for the Decision Tree model on adult1.csv [test] (proportions)

       Predicted
Actual  <=50K >50K Error
  <=50K  71.7  4.4   5.8
  >50K   11.9 12.0  49.7

Overall error: 16.3%, Averaged class error: 27.75%

Rattle timestamp: 2019-09-25 14:47:12 SANGU
```

Figure.4.1 *Error matrix of census dataset*

```
Error matrix for the Decision Tree model on bank.csv [validate] (counts):

     Predicted
Actual   no yes Error
   no  5279 185   3.4
   yes  410 304  57.4

Error matrix for the Decision Tree model on bank.csv [validate] (proportions)

     Predicted
Actual   no yes Error
   no  85.4 3.0   3.4
   yes  6.6 4.9  57.4

Overall error: 9.7%, Averaged class error: 30.4%

Rattle timestamp: 2019-09-19 13:47:02 SANGU
```

Figure.4.2 *Error matrix of bank dataset*

ii) Random forest

The overall error rate of random forest in census dataset and bank dataset is 14.1% and 8.3%

```
Error matrix for the Random Forest model on adult1.csv [validate] (co

       Predicted
Actual  <=50K >50K Error
  <=50K  3227  220   6.4
  >50K    427  740  36.6

Error matrix for the Random Forest model on adult1.csv [validate] (propo

       Predicted
Actual  <=50K >50K Error
  <=50K  69.9  4.8   6.4
  >50K    9.3 16.0  36.6

Overall error: 14.1%, Averaged class error: 21.5%

Rattle timestamp: 2019-10-23 08:55:45 SANGU
```

Figure. 4.3 *Error matrix of census dataset*

```
Error matrix for the Random Forest model on bank.csv [test] (counts):

     Predicted
Actual   no yes Error
   no  5308 197   3.6
   yes  314 360  46.6

Error matrix for the Random Forest model on bank.csv [test] (proportions):

     Predicted
Actual   no yes Error
   no  85.9 3.2   3.6
   yes  5.1 5.8  46.6

Overall error: 8.3%, Averaged class error: 25.1%

Rattle timestamp: 2019-09-19 13:55:28 SANGU
```

Figure. 4.4 *Error matrix of bank dataset*

iii) SVM

The overall error rate of SVM in census dataset and bank dataset is 15.7% and 8.84%

```
Error matrix for the SVM model on adult1.csv [test] (counts):

       Predicted
Actual  <=50K >50K Error
  <=50K  3160  227   6.7
  >50K    482  634  43.2

Error matrix for the SVM model on adult1.csv [test] (proportions):

       Predicted
Actual  <=50K >50K Error
  <=50K  70.2  5.0   6.7
  >50K   10.7 14.1  43.2

Overall error: 15.7%, Averaged class error: 24.95%
```

Figure. 4.5 *Error matrix of census dataset*

```
Error matrix for the SVM model on bank.csv [test] (counts):

      Predicted
Actual   no yes Error
   no  5358 147   2.7
   yes  397 277  58.9

Error matrix for the SVM model on bank.csv [test] (proportions):

      Predicted
Actual   no yes Error
   no  86.7 2.4   2.7
   yes  6.4 4.5  58.9

Overall error: 8.8%, Averaged class error: 30.8%

Rattle timestamp: 2019-09-19 13:59:30 SANGU
```

Figure 4.6 *Error matrix of bank dataset*

iv) Neural network model

The overall error rate of neural net in census dataset and bank dataset is 22.8% and 22.5%

```
Error matrix for the Neural Net model on adult1.csv [test] (counts):

       Predicted
Actual  <=50K >50K Error
  <=50K  3379    8   0.2
  >50K   1017   99  91.1

Error matrix for the Neural Net model on adult1.csv [test] (proportions):

       Predicted
Actual  <=50K >50K Error
  <=50K  75.0  0.2   0.2
  >50K   22.6  2.2  91.1

Overall error: 22.8%, Averaged class error: 45.65%
```

Figure.4.7 *Error matrix of census dataset*

```
Error matrix for the Neural Net model on bank.csv [test] (counts):

      Predicted
Actual   no yes Error
   no  5505   0     0
   yes  674   0   100

Error matrix for the Neural Net model on bank.csv [test] (proportions):

      Predicted
Actual   no yes Error
   no  89.1   0     0
   yes 10.9   0   100

Overall error: 10.9%, Averaged class error: 50%

Rattle timestamp: 2019-09-19 14:02:00 SANGU
```

Figure. 4.8 *Error matrix of bank dataset*

5.2 SENSITIVITY AND SPECIFICITY

Sensitivity is observed by the proposition of positive values divided by the no of observation. It is calculated by true positive and false negative.

$$Sensitivity = \frac{No\ of\ true\ positive\ assesments}{No\ \ of\ all\ positive\ assesments}$$

Specificity is calculated from true negative and false positive. It is predictive values of the system compared to the reference results.

$$Specificity = \frac{No\ of\ true\ negative\ assesments}{No\ of\ all\ negative\ assesments}$$

5.3 ACCURACY

Accuracy is calculated by the specificity and sensitivity values. The accuracy among those algorithms has been calculated by error matrix or confusion matrix. Error matrix give true postive, true negative, false positive and false negative values.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is True Positive values, TN is True Negative, FP is False Negative and FN is False Negative

6. RESULT

The comparison of various models such as decision tree, random forest, SVM and neural network are considered in terms of specificity, sensitivity and accuracy using census and bank datasets. The accuracy and sensitivity of random forest is higher when comparing other three algorithms in census dataset. The accuracy of svm is higher when comparing other three algorithms in bank dataset. Both the datasets had different efficiency over the various models. The model efficiency differ for both the algorithms.

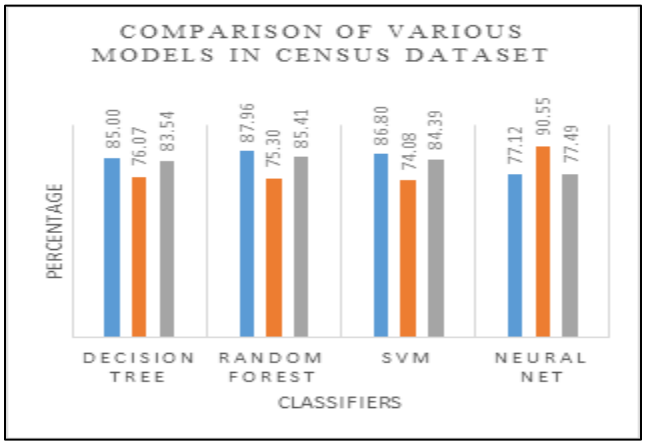

Figure 6.1 *Comparison chart of census dataset*

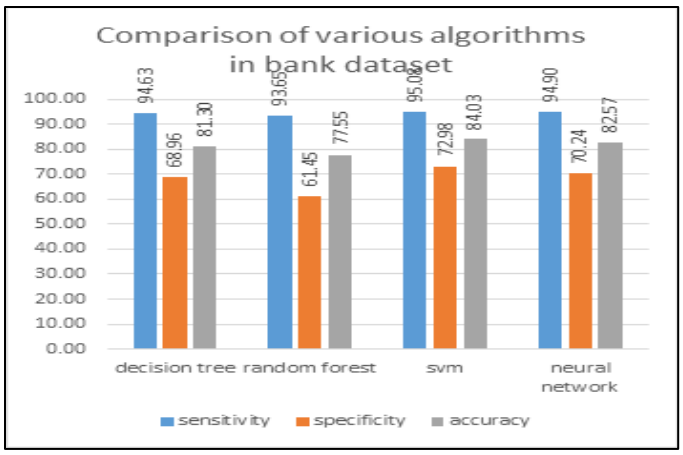

Figure. 6.2 *Comparison chart of bank dataset*

7. CONCLUSIONS

In this project, it compares various models such decision tree, random forest, SVM and neural models in order to get the better accuracy and performance efficiency by referring two datasets. The analysis of the various models shows the efficiency in terms of accuracy, specificity and sensitivity. The accuracy and sensitivity of random forest is higher when comparing other three algorithms in census dataset whereas accuracy of svm is higher in the bank dataset . It is totally different for both datasets. The same algorithms which holds good for both dataset. As the dataset changes the efficiency of the algorithm also varies. In future work, the sensitive features are identified and the selected features are preserved by altering their orignial values with some statistical methods and the performances are analysed with state-of-the art methods.

8 REFERENCES

[1] Dr. Sudhir B. Jagtap, Dr. Kodge B. G, “Census Data Mining and Data Analysis using WEKA”, International Conference in “Emerging Trends in Science, Technology and Management. 2013;10”.

[2] A comparative analysis of classification algorithms in datamining for accuracy, speed and robustness “Dogan, N. Technol M (2013) 14:105”.

[3] S. Archana1, Dr. K. Elangovan, “Survey of Classification Techniques in Data Mining”, International Journal of Computer Science and Mobile Applications.2014;2(2):11.

[4] Dr. A. Bharathi, E. Deepan kumar ,” Survey on Classification Techniques in Data Mining”, International Journal on Recent and Innovation Trends in Computing and Communication. 2014;2(7)

[5] Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study,” Decision Support Systems (Elsevier). 2004;37:543– 558.

[6] “A Comparative Study of Classification Techniques On Adult Data Set” S.Deepajothi , Dr.S.Selvarajan Chettinad college of Engineering and Technology ,TamilNadu,India

[7] “UCI Repository of Machine Learning Databases” by D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, Available at www.ics.uci.edu/~ learn/MLRepository.html, University of California, Irvine, 1998.

[8] Jiawei Han “Datamining: Concepts and Techniques” Second edition, Morgan

[9] N.Sangavi, currently pursuing Post Graduation in Department of Computer Science and Engineering in Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, TamilNadu. She received B.E degree in SNS College of Technology(Autonomous), Coimbatore. She published 5 research papers in International Conferences. Her area of interest is data mining and data analytics.

[10] B.Vinothini, currently Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology (Autonomous), Sathyamangalam, Erode, TamilNadu. She received her B.E in KSR College of Engineering and M.E in SNS College of Technology(Autonomous), Coimbatore. She published 5 research papers in International Conferences and 8 papers in International journals. Her area of interest is data mining and cloud computing

[11] Dr. K.Premalatha, currently Professor and Head of Computer Science and Engineering, Bannari Amman Institute of Technology (Autonomous), Sathyamangalam, Erode, TamilNadu. She received her B.E in IRTT Institute and M.E in Kongu Engineering College, Erode. She had nearly 20 years of experience. She served and held many academic positions and as a PG coordinator in Bannari Amman Institute of Technology. She nearly guided 40+ Ph.D students and doing some funded projects in DRDO and DST- NRDMS projects. She has been effectively published 96 journals in IEEE, Springler and Inderscience publishers and 80 Conference papers. Her area of interest is data mining, machine learning and data analytics.